

Segregation Analysis for Complex Modes of Inheritance

REGINA C. ELANDT-JOHNSON¹

INTRODUCTION

In studies of human traits, observations are often taken on each member of a sibship (called simply a family), and several families are examined in a study. Statistical methods of handling such data are known as segregation analysis.

In this paper, individuals possessing a trait under consideration will be called "affected"; otherwise they will be called "normal."

Let θ be the probability that a child born in a family of certain mating type will be affected. Thus, the probability that in a family of size s there will be r affected children is

$$\Pr(r) = \binom{s}{r} \theta^r (1 - \theta)^{s-r} \quad r = 0, 1, \dots, s. \quad (1)$$

The probability function will be called the *segregation distribution*, and θ will be called the *segregation parameter*.

Consider three parental phenotype mating types: Normal \times Normal, Normal \times Affected, and Affected \times Affected. The early methods developed in segregation analysis apply to the cases with only one segregation distribution within a given parental phenotype mating type. Usually this takes place when a trait is a simple recessive or a dominant with a very low gene frequency (a *rare* trait). However, if a trait is a *common* dominant (i.e., the frequency of the gene A is not too low) or the mode of inheritance is complicated by other parameters such as incomplete penetrance or differential viability or can be controlled by two or more loci, more than one kind of mating is distinguished, each with different segregation distribution within a given phenotype mating type. These matings shall be called *segregation patterns*. Suppose, for example, that a trait is a single, common dominant. Within a phenotype mating type Dominant \times Dominant, there are genotype mating types: $AA \times AA$ and $AA \times Aa$ (both belonging to a segregation pattern with $\theta_1 = 1$) and $Aa \times Aa$ (with $\theta_2 = \frac{3}{4}$).

Statistical methods which take into account different segregation patterns within a given phenotype mating type will be called *complex segregation analysis*. A great advance in this field has been effected by Morton. He and his co-workers (see Chung et al. 1959; Morton 1959, 1962, 1963; Barrai et al. 1965; Dewey et al. 1965; Mi 1967)

Received March 31, 1969; revised July 30, 1969.

This work has been supported by U.S. Public Health Service research grant AI 07975 from the National Institutes of Health.

¹ Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514.

© 1970 by the American Society of Human Genetics. All rights reserved.

derived models which apply when the proportion of sporadic cases, inbreeding coefficient, and prevalence are introduced, in addition to the segregation parameter θ . Peritz (1966) discussed models in which the reproductive behavior of parental genotypes is taken into account.

The objective of this paper is to present a general-purpose model for complex segregation analysis under fairly general conditions. The models derived by Morton and his co-workers utilize the same ideas, and it can be shown that their models can be considered as special cases of the general-purpose model which will be discussed in the next section. Two models, one for incomplete penetrance and one for double common recessives, will be given to demonstrate the use of the general model.

GENERAL-PURPOSE MODEL IN COMPLEX SEGREGATION ANALYSIS

Let H_0 be a certain hypothesis about the mode of inheritance of a trait. Two phenotypic classes are distinguished: affected and normal. Consider here the case when the ascertainment is *through the affected children*, with constant ascertainment probability π ($0 < \pi \leq 1$), so that the distributions of affected children in segregating families are *truncated* binomial distributions. In constructing a model for H_0 , the following steps will be useful:

1. For a given parental *phenotype* mating type, write down all the genotype mating types which can potentially produce affected children. Assuming that the population is in equilibrium, evaluate their expected proportions.
2. Group the genotype mating types according to their segregation patterns. Let m be the number of *distinct* segregation patterns, and θ_t , the segregation parameter for the t th pattern ($t = 1, 2, \dots, m$).
3. Let P_{trs} be the probability that a family of size s within the t th segregation pattern has r affected children and is ascertained. Thus

$$P_{trs} = [1 - (1 - \pi)^r] \binom{s}{r} \theta_t^r (1 - \theta_t)^{s-r}, \quad (2)$$

$t = 1, 2, \dots, m; r = 1, 2, \dots, s; s = 1, 2, \dots, S$. The probability Q_{ts} that such a family has been ascertained at least once is

$$Q_{ts} = \sum_{r=1}^s P_{trs} = 1 - (1 - \pi \theta_t)^s. \quad (3)$$

4. Let ϕ_t be the expected proportion of families within the t th segregation pattern, with

$$\sum_{t=1}^m \phi_t = 1. \quad (4)$$

5. Let P_{rs} be the probability that an ascertained family of size s has r affected children. Thus,

$$P_{rs} = \left(\sum_{t=1}^m \phi_t P_{trs} \right) / \left(\sum_{t=1}^m \phi_t Q_{ts} \right). \quad (5)$$

The probability in (5) is a weighted sum of the proportions P_{trs} . Notice that for $m = 1$ the probability in (5) takes the known form

$$P_{rs} = \frac{[1 - (1 - \pi)^r] \binom{s}{r} \theta^r (1 - \theta)^{s-r}}{1 - (1 - \pi\theta)^s}. \quad (6)$$

6. In the general case, the segregation probability, θ_t , might be a function of some other parameters such as penetrance, different viabilities of gametes, etc. Suppose that there are g *independent* parameters, $\mathfrak{G}' = (\beta_1, \beta_2, \dots, \beta_g)$, so that $\theta_t = \theta_t(\mathfrak{G}')$, $t = 1, 2, \dots, m$. Also, the proportions ϕ_t might be functions of the g parameters $\mathfrak{G}' = (\beta_1, \beta_2, \dots, \beta_g)$ mentioned before, and additionally a function of f *independent* parameters $\mathfrak{A}' = (\alpha_1, \alpha_2, \dots, \alpha_f)$ (for instance, the α 's can be gene frequencies). Let us write the vector $(\mathfrak{A}', \mathfrak{G}')$ in the convenient form

$$(\mathfrak{A}', \mathfrak{G}') = (\alpha_1, \dots, \alpha_f, \beta_1, \dots, \beta_g) = \Upsilon' = (\gamma_1, \gamma_2, \dots, \gamma_M), \quad (7)$$

with $M = f + g$. Thus, in the general case, the probabilities P_{rs} , defined in (5), can be functions of M independent parameters $\Upsilon' = (\gamma_1, \gamma_2, \dots, \gamma_M)$, that is,

$$P_{rs} = P_{rs}(\Upsilon'). \quad (8)$$

7. Let n_s be the number of families, each of size s , ascertained for a given phenotypic mating type, and let a_{rs} be the observed number of families with r affected children, with

$$\sum_{r=1}^s a_{rs} = n_s. \quad (9)$$

For given n_s , the quantities a_{rs} can be considered as random variables having a multinomial distribution with parameters $P_{rs}(\Upsilon')$, respectively, that is,

$$\Pr(a_{1s}, \dots, a_{ss}; \Upsilon') = n_s! \prod_{r=1}^s \left(\frac{P_{rs}^{\alpha_{rs}}}{a_{rs}!} \right). \quad (10)$$

Notice that

$$E(a_{rs}) = n_s P_{rs}, \quad r = 1, 2, \dots, s. \quad (11)$$

8. For particular ascertainment: (a) For truncated complete ($\pi = 1$) all formulas are valid when substituting $\pi = 1$. (b) For single ascertainment ($\pi \rightarrow 0$), use the approximation

$$1 - (1 - \pi)^r \doteq r\pi; \quad 1 - (1 - \pi\theta_t)^s \doteq s\pi\theta_t = Q_{ts}. \quad (12)$$

Substituting (12) into (5) and denoting

$$P'_{trs} = \binom{s-1}{r-1} \theta_t^{r-1} (1 - \theta_t)^{s-r}, \quad (13)$$

it can be shown that (5) takes the form

$$P_{rs} = P'_{rs} = \left(\sum_{t=1}^m \phi_t P'_{trs} \right) / \left(\sum_{t=1}^m \phi_t \theta_t \right). \quad (14)$$

9. The models can be generalized to situations in which n_s is also a random variable, and ascertainment is through the parents.

LIKELIHOOD EQUATIONS AND INFORMATION MATRIX

Denote

$$A = A_{rs}(\gamma) = \sum_{t=1}^m \phi_t P_{trs} \quad \text{and} \quad B = B_{rs}(\gamma) = \sum_{t=1}^m \phi_t Q_{ts}. \quad (15)$$

The probability in (5) can be written in the more convenient form

$$P_{rs} = P_{rs}(\gamma) = \frac{A_{rs}(\gamma)}{B_{rs}(\gamma)} = \frac{A}{B}. \quad (16)$$

The maximum likelihood estimators of the M functionally independent parameters $\gamma' = (\gamma_1, \gamma_2, \dots, \gamma_M)$ are to be found.

For a given sample of n_s families, each of size s , the likelihood function denoted by L_s is given by (10). Taking into account (16) the logarithm of L_s is obtained, that is,

$$\log L_s = \text{Const} + \sum_{r=1}^s [\log A - \log B] a_{rs}. \quad (17)$$

The logarithm of the overall likelihood function, L , say, is

$$\log L = \sum_{s=1}^S \log L_s, \quad (18)$$

where S is the number of different family sizes.

Let $u_{is}(r) = u_{is}(\gamma|r)$ denote the individual score with respect to the parameter γ_i , for given s , in the r th multinomial class. We have

$$u_{is}(r) = u_{is}(\gamma|r) = \frac{1}{A} \frac{\partial A}{\partial \gamma_i} - \frac{1}{B} \frac{\partial B}{\partial \gamma_i}, \quad (19)$$

for $i = 1, 2, \dots, M; r = 1, 2, \dots, s$.

It should be noted that in the general case the derivative $\partial A / \partial \gamma_i$ can be of the form

$$\frac{\partial A}{\partial \gamma_i} = \sum_{t=1}^m \left[\frac{\partial A}{\partial \phi_i} \frac{\partial \phi_t}{\partial \gamma_i} + \frac{\partial A}{\partial P_{trs}} \frac{\partial P_{trs}}{\partial \gamma_i} \right], \quad (20)$$

$i = 1, 2, \dots, M$. The derivative $\partial B / \partial \gamma_i$ has a similar form. The total score for the sample of n_s families, each of size s , U_{is} , say, is

$$U_{is} = \sum_{r=1}^s a_{rs} u_{is}(r), \quad i = 1, 2, \dots, M; \quad (21)$$

and the overall score with respect to the parameter γ_i , U_i , say, is

$$U_i = \sum_{s=1}^S U_{is}, \quad i = 1, 2, \dots, M. \quad (22)$$

To obtain the maximum likelihood estimates, solve M (linearly independent) equations of the form

$$U_i = \sum_{s=1}^S \sum_{r=1}^s \left[\frac{1}{A} \frac{\partial A}{\partial \gamma_i} - \frac{1}{B} \frac{\partial B}{\partial \gamma_i} \right] = 0, \quad (23)$$

$i = 1, 2, \dots, M$.

A unique set of solutions exists provided

$$M < \sum_{s=1}^M (s - 1),$$

which is usually the case.

The elements of the expected information matrix, I_s , are

$$\begin{aligned} I_{iis} &= \sum_{r=1}^s E(a_{rs}) u_{is}^2(r), \quad i = 1, 2, \dots, M; \\ I_{ijs} &= \sum_{r=1}^s E(a_{rs}) u_{is}(r) u_{js}(r), \quad i \neq j, \quad i, j = 1, 2, \dots, M. \end{aligned} \quad (24)$$

The elements of the overall information matrix, I , are

$$I_{ii} = \sum_{s=1}^S I_{iis}; \quad I_{ij} = \sum_{s=1}^S I_{ijs}, \quad i, j = 1, 2, \dots, M, \quad (25)$$

so that

$$I = \sum_{s=1}^S I_s. \quad (26)$$

TESTING HYPOTHESES

1. Suppose that all M parameters are *specified* by the null hypothesis $H_0: \gamma = \gamma_0$. Let

$$U'_s(\gamma_0) = [U_{1s}(\gamma_0), U_{2s}(\gamma_0), \dots, U_{ss}(\gamma_0)] \quad (27)$$

be the vector of scores U_{is} , defined in (21), and evaluated at the point $\gamma = \gamma_0$, and $I_s(\gamma_0)$ be the expected information matrix. Thus the statistic

$$X_{\text{Total}}^2 = \sum_{s=2}^S U'_s(\gamma_0) I_s^{-1}(\gamma_0) U_s(\gamma_0) \quad (28)$$

is approximately distributed as χ^2 with $\frac{1}{2}M(2S - M - 1)$ degrees of freedom.* Notice that the summation begins with $s = 2$, since the families with $s = 1$ child contribute nothing to X_{Total}^2 . The statistic X_{Total}^2 can be used in testing whether the data as a whole fit the hypothesis $H_0: \gamma = \gamma_0$.

* When $s \leq M$, the rank of the matrix I_s is equal to $s - 1$, so that for $s \leq M$ each term $U'_s I_s^{-1} U_s$ in (28) contributes only $(s - 1)$ degrees of freedom. On the other hand, for $s > M$, each of the remaining $(S - M)$ terms contributes M degrees of freedom. Hence the total number of degrees of freedom is $1 + 2 + 3 + \dots + (M - 1) + (S - M)M = [M(M - 1)]/2 + M(S - M) = \frac{1}{2}M(2S - M - 1)$. Notice, of course, that when families of certain sizes $s \leq M$ are missing, the formula for the number of degrees of freedom should be appropriately modified.

If X_{Total}^2 is significant, a further splitting up of the sum (28) would be useful. Let

$$U(\gamma_0) = \sum_{s=1}^S U_s(\gamma_0) \quad \text{and} \quad I(\gamma_0) = \sum_{s=1}^S I_s(\gamma_0) \quad (29)$$

be the overall vector of scores and the expected information matrix of *combined* data from families of all sizes. The statistic

$$X_{\text{Comb}}^2 = U'(\gamma_0) I^{-1}(\gamma_0) U(\gamma_0) \quad (30)$$

is approximately distributed as χ^2 with M degrees of freedom.

Finally the statistic

$$X_{\text{Diff}}^2 = X_{\text{Total}}^2 - X_{\text{Comb}}^2 \quad (31)$$

is approximately distributed as χ^2 with $\frac{1}{2}M(2S - M - 1) - M = \frac{1}{2}M(2S - M - 3)$ degrees of freedom, and is used as a test criterion for heterogeneity.

2. If *all* M parameters are *not specified*, substitute into (28) the pooled estimates, $\hat{\gamma}' = (\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_M)$, say, obtained by solving system (23). Thus the statistic

$$X_{\text{Total}}^{2'} = X_{\text{Diff}}^{2'} = \sum_{s=1}^S U'_s(\hat{\gamma}) I_s^{-1}(\hat{\gamma}) U_s(\hat{\gamma}) \quad (32)$$

is approximately distributed as χ^2 with $\frac{1}{2}M(2S - M - 3)$ degrees of freedom.

3. If among M parameters only K are *specified*, then the statistics (28) and (30) can be evaluated for K of the γ_0 's and $M - K$ of the $\hat{\gamma}$'s. The degrees of freedom have to be appropriately decreased by subtracting $(M - K)$ from the degrees of freedom given in (1).

Some applications are now demonstrated.

SEGREGATION ANALYSIS FOR A SINGLE RECESSIVE WITH INCOMPLETE PENETRANCE

If a genotype aa expresses itself as if it were AA (or Aa), the gene a is said to be incompletely penetrant in a recessive. A similar situation may occur with the gene A , which might be not penetrant in the heterozygote Aa . This can also be extended to traits controlled by more than one locus. The expected proportion of phenotypes which do, in fact, express themselves is called the *penetrance parameter*, denoted by β .

Incomplete penetrance is a complicated phenomenon. It probably depends on the interaction of allelic (or even nonallelic) genes, but environmental conditions may play quite a significant role. It can also be controlled by a separate locus, either linked or independent. Thus this problem cannot be solved uniquely, and different segregation models have to be constructed for different assumptions.

A model is presented here for which these assumptions are rather simplified.

Let A and a be dominant and recessive alleles, respectively, at a single locus. Assume that: (1) the phenotypic expression of genotype aa is incompletely penetrant; (2) penetrance is associated with the locus under consideration or closely linked to it; (3) the environment has very little (random) effect and can be neglected; and (4) there is no physiological effect of genotypes or age of the parents. If these assumptions do not hold, at least approximately, the model will be not valid.

Table 1 exhibits segregation models in segregating families and is self-explanatory. The segregation analysis for this model is as follows:

PARENTAL PHENOTYPE MATINGS NORMAL \times AFFECTED

Using the notations introduced in the second section,

$$\begin{aligned}\phi_1 &= h = q(1 - \beta)/[2p + q(1 - \beta)] ; \\ \phi_2 &= 1 - h = 2p/[2p + q(1 - \beta)] ,\end{aligned}\tag{33}$$

TABLE 1
SEGREGATION MODEL FOR SIMPLE AUTOSOMAL RECESSIVE WITH
INCOMPLETE PENETRANCE (SEGREGATING FAMILIES)

MATINGS	EXPECTED PROPORTIONS OF MATINGS WHICH CAN PRODUCE AFFECTED CHILDREN		SEGREGA- TION PA- RAMETER (θ_t)	SEGREGATION DISTRIBUTION
	In a Population	In a Given Parental Mating Type (ϕ_t)		
One Parent Unaffected				
$aa \times aa \dots\dots\dots$	$2q^4\beta(1-\beta)$	$\frac{q(1-\beta)}{2p+q(1-\beta)}$ $= h$	β	$\binom{s}{r}\beta^r(1-\beta)^{s-r}$
$aa \times Aa \dots\dots\dots$	$4pq^3\beta$	$\frac{2p}{2p+q(1-\beta)}$ $= 1-h$	$\frac{1}{2}\beta$	$\binom{s}{r}(\frac{1}{2}\beta)^r(1-\frac{1}{2}\beta)^{s-r}$
Total $\dots\dots\dots$	$2q^3\beta[2p+q(1-\beta)]$	1	$\dots\dots\dots$	$\dots\dots\dots$
Both Parents Unaffected				
$aa \times aa \dots\dots\dots$	$q^4(1-\beta)^2$	$\frac{q^2(1-\beta)^2}{[2p+q(1-\beta)]^2}$ $= h^2$	β	$\binom{s}{r}\beta^r(1-\beta)^{s-r}$
$aa \times Aa \dots\dots\dots$	$4pq^3(1-\beta)$	$\frac{4pq(1-\beta)}{[2p+q(1-\beta)]^2}$ $= 2h(1-h)$	$\frac{1}{2}\beta$	$\binom{s}{r}(\frac{1}{2}\beta)^r(1-\frac{1}{2}\beta)^{s-r}$
$Aa \times Aa \dots\dots\dots$	$4p^2q^2$	$\frac{4p^2}{[2p+q(1-\beta)]^2}$ $= (1-h)^2$	$\frac{1}{4}\beta$	$\binom{s}{r}(\frac{1}{4}\beta)^r(1-\frac{1}{4}\beta)^{s-r}$
Total $\dots\dots\dots$	$q^2[2p+q(1-\beta)]^2$	1	$\dots\dots\dots$	$\dots\dots\dots$

where $p = 1 - q$. Further, $\theta_1 = \beta$ and $\theta_2 = \frac{1}{2}\beta$. Hence

$$P_{1rs} = [1 - (1 - \pi)^r] \binom{s}{r} \beta^r (1 - \beta)^{s-r}; \quad (34)$$

$$P_{2rs} = [1 - (1 - \pi)^r] \binom{s}{r} (\frac{1}{2}\beta)^r (1 - \frac{1}{2}\beta)^{s-r},$$

and

$$Q_{1s} = 1 - (1 - \pi\beta)^s, \quad Q_{2s} = 1 - (1 - \frac{1}{2}\pi\beta)^s. \quad (35)$$

Finally,

$$P_{rs} = P_{rs}(\beta, q) = \frac{hP_{1rs} + (1 - h)P_{2rs}}{hQ_{1s} + (1 - h)Q_{2s}}. \quad (36)$$

The particular cases, $\pi = 1$ and $\pi \rightarrow 0$, are straightforward.

PARENTAL PHENOTYPE MATINGS NORMAL \times NORMAL

The second (bottom) part of table 1 is self-explanatory. The coefficients ϕ_1, ϕ_2, ϕ_3 are $h^2, 2h(1 - h), (1 - h)^2$, respectively. The probabilities P_{1rs}, P_{2rs} and Q_{1s}, Q_{2s} are the same as in (34) and (35), respectively. Additionally,

$$P_{3rs} = [1 - (1 - \pi)^r] \binom{s}{r} (\frac{1}{4}\beta)^r (1 - \frac{1}{4}\beta)^{s-r}, \quad (37)$$

and

$$Q_{3s} = 1 - (1 - \frac{1}{4}\beta)^s. \quad (38)$$

The multinomial parameters, P_{rs} , are

$$P_{rs} = P_{rs}(\beta, q) = \frac{h^2P_{1rs} + 2h(1 - h)P_{2rs} + (1 - h)^2P_{3rs}}{h^2Q_{1s} + 2h(1 - h)Q_{2s} + (1 - h)^2Q_{3s}}. \quad (39)$$

The logarithms of the likelihood functions, $L_s(\beta, q)$ and $L(\beta, q)$, for both parental phenotype mating types are straightforward to evaluate, using the results of section 3 (for details, see Appendix). This model is illustrated by the following example:

Example 1

Rheumatic fever is a certain acute form of rheumatoid arthritis. There is some doubt whether heredity plays an important role in this disease (O'Brien et al. 1965). However, some investigators believe that heredity plays some role, and it is usually suggested that it might be a single autosomal recessive.

Table 2 presents some data from Stevenson and Cheeseman (1953) on rheumatic fever in children from three to 18 years old in 388 families with both parents unaffected, each family ascertained by one proband.

Although one may assume that the ascertainment was single, he also should notice that there are only a few families with more than two affected children, and the assumption $\pi \rightarrow 0$ might be incorrect. Since π is unknown, the analysis is presented for two limiting cases: $\pi = 1$ and $\pi \rightarrow 0$.

The analysis was performed using an IBM 360/75 computer. The program in FORTRAN IV was written by Mrs. Ellen Kaplan. The following results were obtained:

a) *Complete (truncated) ascertainment* ($\pi = 1$). The *ML*-estimates are: $\hat{\beta} = 0.2339$, $\hat{q} = 0.3378$, and the estimated variance-covariance matrix is

$$I^{-1}(\beta, \hat{q}) \doteq V(\beta, \hat{q}) = \begin{bmatrix} 1.3015 & -0.4654 \\ -0.4654 & 0.1675 \end{bmatrix}.$$

The $X_{\text{Total}}^2 = X_{\text{Diff}}^2 = 36.81$ with 21 df. It is not significant at the significance level $\alpha = .01$, but it is significant at $\alpha = .05$.

TABLE 2

JUVENILE RHEUMATISM IN 388 FAMILIES RECORDED BY ONE AFFECTED CHILD

FAMILY SIZE s	NO. OF FAMILIES n_s	NO. OF ALL CHIL- DREN sn_s	NO. OF FAMILIES WITH r AFFECTED CHILDREN a_{rs}					TOTAL NO. OF AFFECTED CHILDREN $\sum_{r=1}^s r a_{rs}$
			No. of Affected Children (r)					
			1	2	3	4	5	
1.....	19	19	19	19
2.....	67	134	66	1	68
3.....	73	219	66	5	2	82
4.....	72	288	63	9	81
5.....	46	230	38	8	54
6.....	43	258	33	8	2	55
7.....	23	161	19	4	27
8.....	21	168	14	6	1	29
9.....	13	117	8	4	1	19
10.....	8	80	2	3	1	1	1	20
11.....	1	11	1	1
12.....	1	12	1	1
13.....	1	13	1	1
Total....	388	1,710	331	48	7	1	1	457

NOTE.—Both parents unaffected.

SOURCE.—Stevenson and Cheeseman (1953).

b) *Single ascertainment* ($\pi \rightarrow 0$). Here, $\hat{\beta} = 0.124$, $\hat{q} = 0.3310$ and

$$I^{-1}(\beta, \hat{q}) \doteq V(\beta, \hat{q}) = \begin{bmatrix} 5.8353 & -1.0945 \\ -1.0945 & 0.2055 \end{bmatrix}.$$

Then $X_{\text{Total}}^2 = X_{\text{Diff}}^2 = 30.56$ with 20 df. (Note that in the case of single ascertainment, s has to be replaced by $s - 1$ so that the number of degrees of freedom is $[(M - 1)(M - 2)]/2 + M(S - M - 1)$. Here X_{Total}^2 is not significant at $\alpha = .05$.)

The data "almost" fit the hypothesis of a single recessive with incomplete penetrance, although the fit is not too obvious. It would be useful to have more (medical) information.

A MODEL OF SEGREGATION ANALYSIS FOR A COMMON AUTOSOMAL DOUBLE RECESSIVE

It is suspected that a trait might be a double recessive, *aabb*, and the frequencies of the genes *a* and *b* are both very low (i.e., the trait is rare). The parental matings

TABLE 3
SEGREGATION MODEL FOR (COMMON) DOUBLE AUTOSOMAL
RECESSIVE (SEGREGATING FAMILIES)

MATINGS	EXPECTED PROPORTIONS OF MATINGS WHICH CAN PRODUCE AFFECTED CHILDREN		SEGREGA- TION PA- RAMETER (θ_t) (UNDER H_0)	SEGREGATION DISTRIBUTION (UNDER H_0)
	In a Population	In a Given Parental Mating Type (ϕ_t)		
One Parent Unaffected				
$Aabb \times aabb \dots\dots$	$4p_1q_1^3q_2^4$	$\frac{p_1q_2}{p_1+p_2} \left. \vphantom{\frac{p_1q_2}{p_1+p_2}} \right\} 1-h$	$\frac{1}{2}$	$\binom{s}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{s-r}$
$aaBb \times aabb \dots\dots$	$4q_1^4p_2q_2^3$	$\frac{p_2q_1}{p_1+p_2}$	$\frac{1}{2}$	$\binom{s}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{s-r}$
$AaBb \times aabb \dots\dots$	$8p_1q_1^3p_2q_2^3$	$\frac{2p_1p_2}{p_1+p_2} = h$	$\frac{1}{4}$	$\binom{s}{r} \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{s-r}$
Total	$4q_1^3q_2^3(p_1+p_2)$	1
Both Parents Unaffected				
$Aabb \times Aabb \dots\dots$	$4p_1^2q_1^2q_2^4$	$\frac{p_1^2q_2^2}{(p_1+p_2)^2} \left. \vphantom{\frac{p_1^2q_2^2}{(p_1+p_2)^2}} \right\} (1-h)^2$	$\frac{1}{4}$	$\binom{s}{r} \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{s-r}$
$aaBb \times Aabb \dots\dots$	$8p_1q_1^3p_2q_2^3$	$\frac{2p_1q_1p_2q_2}{(p_1+p_2)^2}$	$\frac{1}{4}$	$\binom{s}{r} \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{s-r}$
$aaBb \times aaBb \dots\dots$	$4q_1^4p_2^2q_2^2$	$\frac{q_1^2p_2^2}{(p_1+p_2)^2} \left. \vphantom{\frac{q_1^2p_2^2}{(p_1+p_2)^2}} \right\} 2h(1-h)$	$\frac{1}{4}$	$\binom{s}{r} \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{s-r}$
$AaBb \times Aabb \dots\dots$	$16p_1^2q_1^2p_2q_2^3$	$\frac{q_1^2p_2^2}{(p_1+p_2)^2}$	$\frac{1}{8}$	$\binom{s}{r} \left(\frac{1}{8}\right)^r \left(\frac{7}{8}\right)^{s-r}$
$AaBb \times aaBb \dots\dots$	$16p_1q_1^3p_2^2q_2^2$	$\frac{4p_1^2p_2q_2}{(p_1+p_2)^2}$	$\frac{1}{8}$	$\binom{s}{r} \left(\frac{1}{8}\right)^r \left(\frac{7}{8}\right)^{s-r}$
$AaBb \times AaBb \dots\dots$	$16p_1^2q_1^2p_2^2q_2^2$	$\frac{4p_1q_1p_2^2}{(p_1+p_2)^2} = h^2$	$\frac{1}{16}$	$\binom{s}{r} \left(\frac{1}{16}\right)^r \left(\frac{15}{16}\right)^{s-r}$
Total	$4q_1^2q_2^2(p_1+p_2)^2$	1

Normal \times Affected are likely to be $AaBb \times aabb$, and the parental matings Normal \times Normal are likely to be $AaBb \times AaBb$. However, if a trait is *common*, more than one segregation pattern can occur within a given parental phenotype mating type. Table 3 exhibits all genotype matings, grouped into appropriate segregation patterns. This table is self-explanatory. Under the null hypothesis that the trait is a double recessive, the segregation distributions are given in the last column of table 3

PARENTAL MATINGS NORMAL \times AFFECTED

Let θ be the segregation parameter for the matings Single Heterozygote \times Double Recessive. From table 3, $\theta_1 = \theta$, $\theta_2 = \frac{1}{2}\theta$, and $\phi_1 = 1 - h$, $\phi_2 = h$.

PARENTAL MATINGS NORMAL \times NORMAL

Let θ be the segregation parameter for Single Heterozygote \times Single Heterozygote (see table 3). We have $\theta_1 = \theta$, $\theta_2 = \frac{1}{2}\theta$, $\theta_3 = \frac{1}{4}\theta$, and $\phi_1 = (1 - h)^2$, $\phi_2 = 2h(1 - h)$, $\phi_3 = h^2$.

Suppose that the hypothesis $H_0: \theta = \theta_0$ (i.e., for Normal \times Affected, $H_0: \theta = \frac{1}{2}$; for Normal \times Normal, $H_0: \theta = \frac{1}{4}$) is to be tested. For each mating type, construct the likelihood function $L(\theta, h)$ as described in section 3, find the pooled estimators, $\hat{\theta}$ and \hat{h} , and calculate X_{Total}^2 , X_{Comb}^2 , and X_{Diff}^2 with θ given the specified value, θ_0 , and unspecified h put equal to \hat{h} .

Example 2

Psoriasis is a chronic inflammatory disease of skin characterized by rounded erythematous dry scaling patches of various sizes covered by grayish-white scales. It has been suggested that it might be an autosomal double recessive (Steinberg et al. 1951). The data in table 4 are taken from the paper by Steinberg et al. (1951) and give the distributions of affected patients in 409 families, with both parents unaffected.

Although all patients were probands, the analysis was performed for both limiting cases, $\pi = 1$ and $\pi \rightarrow 0$, for the same reason as in example 1. Using the computer program in FORTRAN IV, the following results were obtained:

a) *Complete ascertainment* ($\pi = 1$). Values obtained were $\hat{\theta} = 0.0991$, $\hat{h} = 0.7617$. The variance-covariance matrix evaluated at $\theta = \theta_0 = 0.25$ and $h = \hat{h} = 0.7617$ is

$$I^{-1}(\theta_0, \hat{h}) \doteq V(\hat{\theta}, \hat{h}) = \begin{bmatrix} 0.6709 & 0.3175 \\ 0.3175 & 0.1508 \end{bmatrix}.$$

Here $S = 13$, $M = 2$, $K = 1$. To test the hypothesis $H_0: \theta = 0.25$, calculations yield $X_{\text{Total}}^2 = 46.32$ with 22 df, $X_{\text{Comb}}^2 = 36.61$ with 1 df, and $X_{\text{Diff}}^2 = 9.71$ with 21 df.

b) *Incomplete ascertainment* ($\pi \rightarrow 0$). Values obtained were $\hat{\theta} = 0.0549$, $\hat{h} = 0.7958$, and

$$I^{-1}(\theta_0, \hat{h}) \doteq V(\hat{\theta}, \hat{h}) = \begin{bmatrix} 0.2710 & 0.1393 \\ 0.1393 & 0.0720 \end{bmatrix};$$

$X^2_{\text{Total}} = 115.69$ with 21 df, $X^2_{\text{Comb}} = 110.20$ with 1 df, and $X^2_{\text{Diff}} = 5.49$ with 20 df.

It appears from both analyses that the data are homogeneous (X^2_{Diff} in both calculations is not significant), but the hypothesis that psoriasis is a common double recessive trait does not fit the data.

TABLE 4
PSORIASIS IN 409 FAMILIES

FAMILY SIZE s	No. OF FAMILIES n_s	No. OF ALL CHIL- DREN sn_s	No. OF FAMILIES WITH r AFFECTED CHILDREN a_{rs}				TOTAL No. OF AFFECTED CHILDREN $\sum_{r=1}^s r a_{rs}$
			No. of Affected Children (r)				
			1	2	3	4	
1.....	22	22	22	22
2.....	50	100	45	5	55
3.....	72	216	67	5	77
4.....	61	244	55	6	67
5.....	62	310	59	3	65
6.....	37	222	32	5	42
7.....	28	196	26	2	30
8.....	24	192	22	1	1	27
9.....	24	216	22	2	26
10.....	13	130	11	1	1	16
11.....	7	77	5	2	9
12.....	3	36	2	1	4
13.....	6	78	5	1	9
Total.....	409	2,039	373	33	2	1	449

NOTE.—Both parents unaffected.

SOURCE.—Steinberg et al. (1951).

SUMMARY

The paper presents a fairly general model for segregation analysis, when more than one segregation distribution can occur for a given parental phenotype mating type. In this model the expected proportions of affected children in families, each of size s , and within a given phenotype mating type, are functions of segregation parameters weighted by the relative frequencies of different segregation patterns (i.e., matings with distinct segregation distributions within a given phenotype mating). Models for autosomal single recessives with incomplete penetrance and for common autosomal double recessives are derived. Two numerical examples, one for each model, are calculated using a high-speed computer program.

APPENDIX

DERIVATION OF LIKELIHOOD EQUATIONS FOR ESTIMATION OF
PENETRANCE PARAMETER β AND GENE FREQUENCY q

COMPLETE ASCERTAINMENT

First consider the situation when the ascertainment is complete, that is, $\pi = 1$.

From formula (39), $h = q(1 - \beta)/[2 - q(1 + \beta)]$. Thus,

$$\frac{\partial h}{\partial q} = \frac{2(1 - \beta)}{[2 - q(1 + \beta)]^2}, \quad \frac{\partial h}{\partial \beta} = -\frac{2q(1 - q)}{[2 - q(1 + \beta)]^2}. \quad (i)$$

Let

$$X_1 = \binom{s}{r} \beta^r (1 - \beta)^{s-r}, \quad X_2 = \binom{s}{r} (\tfrac{1}{2}\beta)^r (1 - \tfrac{1}{2}\beta)^{s-r},$$

$$X_3 = \binom{s}{r} (\tfrac{1}{4}\beta)^r (1 - \tfrac{1}{4}\beta)^{s-r}.$$

Notice that X_t corresponds to P_{trs} (defined in equations [34] and [37] for $t = 1, 2, 3$), with $\pi = 1$. Let $Y_1 = 1 - (1 - \beta)^s$, $Y_2 = 1 - (1 - \tfrac{1}{2}\beta)^s$, $Y_3 = 1 - (1 - \tfrac{1}{4}\beta)^s$. Also notice that Y_t corresponds to Q_{ts} (defined in equations [35] and [38] for $t = 1, 2, 3$), with $\pi = 1$. Thus,

$$\begin{aligned} A &= h^2 X_1 + 2h(1 - h)X_2 + (1 - h)^2 X_3, \\ B &= h^2 Y_1 + 2h(1 - h)Y_2 + (1 - h)^2 Y_3. \end{aligned} \quad (ii)$$

Hence,

$$\begin{aligned} \frac{\partial A}{\partial q} &= \frac{\partial A}{\partial h} \cdot \frac{\partial h}{\partial q} = 2[(X_2 - X_3) + h(X_1 - 2X_2 + X_3)] \frac{\partial h}{\partial q}, \\ \frac{\partial B}{\partial q} &= \frac{\partial B}{\partial h} \cdot \frac{\partial h}{\partial q} = 2[(Y_2 - Y_3) + h(Y_1 - 2Y_2 + Y_3)] \frac{\partial h}{\partial q}, \end{aligned} \quad (iii)$$

and

$$u_{s;q} = \frac{1}{A} \frac{\partial A}{\partial q} - \frac{1}{B} \frac{\partial B}{\partial q}. \quad (iv)$$

Also,

$$\begin{aligned} \frac{\partial X_1}{\partial \beta} &= \binom{s}{r} [r\beta^{r-1}(1 - \beta) - (s - r)\beta^r(1 - \beta)^{s-r-1}] \\ &= \binom{s}{r} \beta^r (1 - \beta)^{s-r} \left(\frac{r}{\beta} - \frac{s - r}{1 - \beta} \right). \end{aligned}$$

Putting

$$Z_1 = \frac{r}{\beta} - \frac{s - r}{1 - \beta}$$

yields

$$\frac{\partial X_1}{\partial \beta} = X_1 Z_1.$$

Similarly putting

$$Z_2 = \tfrac{1}{2} \left(\frac{r}{\tfrac{1}{2}\beta} - \frac{s - r}{1 - \tfrac{1}{2}\beta} \right) \quad \text{and} \quad Z_3 = \tfrac{1}{4} \left(\frac{r}{\tfrac{1}{4}\beta} - \frac{s - r}{1 - \tfrac{1}{4}\beta} \right)$$

yields

$$\frac{\partial X_2}{\partial \beta} = X_2 Z_2 \quad \text{and} \quad \frac{\partial X_3}{\partial \beta} = X_3 Z_3 .$$

Analogously

$$\frac{\partial Y_1}{\partial \beta} = s(1 - \beta)^{s-1} = W_1, \quad \frac{\partial Y_2}{\partial \beta} = \frac{1}{2}s(1 - \frac{1}{2}\beta)^{s-1} = W_2,$$

$$\frac{\partial Y_3}{\partial \beta} = \frac{1}{4}s(1 - \frac{1}{4}\beta)^{s-1} = W_3 .$$

Therefore,

$$\begin{aligned} \frac{\partial A}{\partial \beta} &= \frac{\partial A}{\partial h} \cdot \frac{\partial h}{\partial \beta} + \sum_{i=1}^3 \frac{\partial A}{\partial X_i} \cdot \frac{\partial X_i}{\partial \beta} = 2[(X_2 - X_3) + h(X_1 - 2X_2 + X_3)] \frac{\partial h}{\partial \beta} \\ &\quad + [h^2 X_1 Z_1 + 2h(1 - h)X_2 Z_2 + (1 - h)^2 X_3 Z_3], \end{aligned} \quad (\text{v})$$

and

$$\begin{aligned} \frac{\partial B}{\partial \beta} &= \frac{\partial B}{\partial h} \frac{\partial h}{\partial \beta} + \sum_{i=1}^3 \frac{\partial B}{\partial X_i} \frac{\partial X_i}{\partial \beta} = 2[(Y_2 - Y_3) + h(Y_1 - 2Y_2 + Y_3)] \frac{\partial h}{\partial \beta} \\ &\quad + [h^2 W_1 + 2h(1 - h)W_2 + (1 - h)^2 W_3]. \end{aligned}$$

Also,

$$u_{\beta s} = \frac{1}{A} \frac{\partial A}{\partial \beta} - \frac{1}{B} \frac{\partial B}{\partial \beta} \quad \text{and} \quad u_{qs} = \frac{1}{A} \frac{\partial A}{\partial q} - \frac{1}{B} \frac{\partial B}{\partial q}. \quad (\text{vi})$$

The likelihood equations are

$$\begin{aligned} \sum_{s=1}^S \sum_{r=1}^s \frac{1}{A} \frac{\partial A}{\partial q} a_{rs} - \sum_{s=1}^S \frac{1}{B} \frac{\partial B}{\partial q} n_s &= 0, \\ \sum_{s=1}^S \sum \frac{1}{A} \frac{\partial A}{\partial \beta} a_{rs} - \sum_{s=1}^S \frac{1}{B} \frac{\partial B}{\partial \beta} n_s &= 0. \end{aligned} \quad (\text{vii})$$

SINGLE INCOMPLETE ASCERTAINMENT ($\pi \rightarrow 0$)

Now put

$$\begin{aligned} X_1 &= \binom{s-1}{r-1} \beta^{r-1} (1 - \beta)^{s-r}, \quad X_2 = \binom{s-1}{r-1} (\frac{1}{2}\beta)^{r-1} (1 - \frac{1}{2}\beta)^{s-r}, \\ X_3 &= \binom{s-1}{r-1} (\frac{1}{4}\beta)^{r-1} (1 - \frac{1}{4}\beta)^{s-r}, \end{aligned}$$

where X_i corresponds to $P'_{i,s}$ given in (18). Putting $Y_1 = Y_2 = Y_3 = 1$ (after a simple modification of equation [14]),

$$\begin{aligned} A &= h^2 X_1 + h(1 - h)X_2 + \frac{1}{4}(1 - h)^2 X_3, \\ B &= h^2 Y_1 + h(1 - h)Y_2 + \frac{1}{4}(1 - h)^2 Y_3 = \frac{1}{4}(1 + h)^2. \end{aligned} \quad (\text{viii})$$

Analogous to complete ascertainment,

$$\begin{aligned}\frac{\partial A}{\partial q} &= \frac{1}{2}[(2X_2 - X_3) + h(4X_1 - 4X_2 + X_3)] \frac{\partial h}{\partial q}, \\ \frac{\partial B}{\partial q} &= \frac{1}{2}[(2Y_2 - Y_3) + h(4Y_1 - 4Y_2 + Y_3)] \frac{\partial h}{\partial q} = \frac{1}{2}(1 + h) \frac{\partial h}{\partial q},\end{aligned}\quad (1x)$$

where $\partial h/\partial q$ is given in formula (i). Putting

$$\begin{aligned}Z_1 &= \frac{r-1}{\beta} - \frac{s-r}{1-\beta}, \quad Z_2 = \frac{1}{2} \left[\frac{r-1}{\frac{1}{2}\beta} - \frac{s-r}{1-\frac{1}{2}\beta} \right], \\ Z_3 &= \frac{1}{4} \left[\frac{r-1}{\frac{1}{4}\beta} - \frac{s-r}{1-\frac{1}{4}\beta} \right]\end{aligned}$$

yields

$$\frac{\partial X_1}{\partial \beta} = X_1 Z_1, \quad \frac{\partial X_2}{\partial \beta} = X_2 Z_2, \quad \frac{\partial X_3}{\partial \beta} = X_3 Z_3.$$

Also,

$$\frac{\partial Y_1}{\partial \beta} = \frac{\partial Y_2}{\partial \beta} = \frac{\partial Y_3}{\partial \beta} = 0.$$

Hence,

$$\begin{aligned}\frac{\partial A}{\partial \beta} &= \frac{1}{2}[(2X_2 - X_3) + h(4X_1 - 4X_2 + X_3)] \frac{\partial h}{\partial \beta} \\ &+ [h^2 X_1 Z_1 + h(1-h)X_2 Z_2 + \frac{1}{4}(1-h)^2 X_3 Z_3],\end{aligned}\quad (x)$$

and

$$\frac{\partial B}{\partial \beta} = \frac{1}{2}[(2Y_2 - Y_3) + h(4Y_1 - 4Y_2 + Y_3)] \frac{\partial h}{\partial \beta} = \frac{1}{2}(1 + h) \frac{\partial h}{\partial \beta}.$$

The likelihood equations are

$$\begin{aligned}\sum_{s=2}^S \sum_{r=1}^s \frac{1}{A} \frac{\partial A}{\partial q} a_{rs} - \sum_{s=2}^S \frac{1}{B} \frac{\partial B}{\partial q} n_s &= 0, \\ \sum_{s=2}^S \sum_{r=1}^s \frac{1}{A} \frac{\partial A}{\partial \beta} a_{rs} - \sum_{s=2}^S \frac{1}{B} \frac{\partial B}{\partial \beta} n_s &= 0.\end{aligned}\quad (xi)$$

The likelihood equations for the problem of a double recessive can be obtained in a similar manner. Notice that in this problem the parameters h and θ are functionally independent, so that the derivatives $\partial A/\partial h$, $\partial A/\partial \theta$, and $\partial B/\partial h$, $\partial B/\partial \theta$ take a simpler form than in the case of a single recessive with incomplete penetrance.

REFERENCES

- BARRAI, I.; MI, M. P.; MORTON, N. E.; AND YASUDA, N. 1965. Estimation of prevalence under incomplete selection. *Amer. J. Hum. Genet.* **17**:221-236.
 CHUNG, C. S.; ROBINSON, O. W.; and MORTON, N. E. 1959. A note on deaf mutism. *Ann. Hum. Genet.* **23**:357-366.
 DEWEY, W. J.; BARRAI, I.; MORTON, N. E.; and MI, M. P. 1965. Recessive genes in severe mental defect. *Amer. J. Hum. Genet.* **17**:237-256.
 MI, M. P. 1967. Segregation analysis. *Amer. J. Hum. Genet.* **19**:313-321.

- MORTON, N. E. 1959. Genetic tests under incomplete ascertainment. *Amer. J. Hum. Genet.* **11**:1-16.
- MORTON, N. E. 1962. Segregation and linkage. Pp. 17-52 in J. BURDETTE (ed.), *Methodology in human genetics*. Holden-Day, San Francisco.
- MORTON, N. E. 1963. Models and evidence in human population genetics. Pp. 935-951 in *Genetics today*, vol. 3. (Proc. 11th Int. Congr. Genet., The Hague.) Pergamon Press, Oxford.
- O'BRIEN, W. M.; LI, C. C.; and TAYLOR, F. H. 1965. Penetrance and the distribution of sib-pair types, exemplified by taste ability and rheumatoid arthritis. *J. Chronic Dis.* **18**:675-680.
- PERITZ, E. 1966. On some models for segregation analysis. *Ann. Hum. Genet.* **30**:183-192.
- STEINBERG, A. G.; BECKER, S. W.; and FITZPATRICK, T. B. 1951. A genetic and statistical study of psoriasis. *Amer. J. Hum. Genet.* **3**:267-281.
- STEVENSON, A. C., and CHEESEMAN, E. A. 1953. Heredity and rheumatic fever: a study of 462 families ascertained by an affected child and 51 families ascertained by an affected mother. *Ann. Eugen.* **17**:177-211.